Prominence-based evaluation of L2 prosody

Kallio, Heini Henriikka

ISCA
2018

Kallio , H H , Suni , A , Virkkunen , P & Simko , J 2018 , Prominence-based evaluation of L2 prosody . in 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018) : Speech Research for Emerging Markets in Multilingual Societies . Interspeech , ISCA , Baixas , pp. 1838-1842 , Annual Conference of the International Speech Communication Association , Hyderabad , India , 02/09/2018 . https://doi.org/10.21437/Interspe

http://hdl.handle.net/10138/306202
https://doi.org/10.21437/Interspeech.2018-1873

unspecified
publishedVersion

Interspeech 2018
2-6 September 2018, Hyderabad

Prominence-based evaluation of L2 prosody

Heini Kallio, Antti Suni, Päivi Virkkunen, Juraj Šimko

University of Helsinki, Finland
heini.h.kallio@helsinki.fi

Abstract

Prosody in terms of word and sentence stress is one of the most difficult features for many second language (L2) speakers to learn and it can be hypothesized that assessing the learner's prosodic abilities could provide a good measure for assessing the learners' spoken language skills in general. Automatic assessment is, however, dependent on reliable automatic analyses of prosodic features for comparing the productions between native (L1) and L2 speech. Here we investigate, whether estimated prosodic prominence levels of syllables can be used to predict the prosodic competence of Finnish learners of Swedish. Syllable level prominence was estimated for 99 L2 and 25 native Swedish utterances using continuous wavelet transform analysis with combinations of f_0, energy, and duration features. The L2 utterances were assessed by four expert raters using the revised CEFR scale for prosodic features. Correlations of prominence estimates for L2 utterances with estimates for L1 utterances and linguistic stress patterns were used as a measure of prosodic proficiency of the L2 speakers. The results show that these estimates correlate significantly with the assessments of expert raters. Overall, the results provide strong support for the use of the wavelet-based prominence estimation techniques in automatic assessment of L2 proficiency.

Index Terms: L2 proficiency assessments, prosody, wavelets

1. Introduction

Prosodic features are important cues in assessing spoken second or foreign language (L2) proficiency. The minimum requirement and baseline for assessments is intelligibility, which is often compromised more by prosodic than phonetic errors [1]. One important contributing feature has been shown to be realization of lexical [2] and utterance-level [3] stress. A stress-bearing syllable is typically characterized by an increase in f_0, duration and intensity [4, 5, 6, 7]. These signal characteristics combine in a complex and language dependent manner; the overall increase can be quantified in terms of syllable prominence.

In the current study we investigate the effect of syllable level prominence on the perceived proficiency in Finnish learners of Swedish. We present a new automatic method of quantifying syllable-level prominence based on continuous wavelet transform. The prominence estimates for L2 speakers are compared with those realized by native speakers as well as lexical stress patterns. We show that the degree of agreement between the prominence patterns correlates positively with the L2 proficiency as assessed by expert raters.

Appropriate placement of word and sentence stress is important in chunking speech into linguistically relevant units. The speaker's native language (L1) affects both the perception and production of L2 stress patterns [8]. For L2 speakers of English the stress structure of English is found to be challenging [3], and preliminary findings on the L2 Swedish of Finnish students suggest similar implications. Previous research on L2 English have shown that low proficiency L2 speakers' speech contain less stressed words than high proficiency or native speakers' speech [9] and that L2 speakers tend to place equal stress on every word [10]. Moreover, disfluent speech often contains repetitions, corrections and false starts, which may cause emphasizing wrong words or syllables in an utterance.

Finland Swedish (FS) is a variant of Swedish spoken in Finland. Prosody of FS is studied only marginally, but it is believed to be more similar to Finnish than standard Swedish (spoken in Sweden) and differ from standard Swedish with regard to phonemic characteristics as well as prosodic features like realization of sentence and word stress. For example, the lexical pitch accents $acute$ and $grave$ that are characteristic for standard Swedish, are absent in FS. However, the linguistic properties of standard Swedish define also the stress structure of FS; the placement of word stress varies in Swedish, while Finnish has fixed word stress on the initial syllable. Additionally, duration contrast related to linguistic quantity is more consistent in Finnish than in Swedish where it is strongly related to stress [11]. Moreover, native speakers of FS seem to vary their f_0 more than Finnish L2 learners when speaking Swedish [12]. It can thus be presumed, that there are differences in the use of f_0, duration (and intensity) between Finnish and FS prosody, including realization of stress-related prominence patterns. These differences can interfere with the perceived oral proficiency of Finnish L2 learners of FS.

Swedish is a compulsory subject in basic education in Finland, and the national matriculation examination test of L2 Swedish is taken by approximately 8000 upper secondary school students yearly [13], which makes Swedish the second most tested L2 in Finland. The Finnish Ministry of Education and Culture have set a goal to include L2 speaking tests to the final examination of upper secondary education by 2022 [14]. This study is part of a larger research project, where automatic assessment methods are studied and developed for the use of the upcoming large-scale speaking tests [15].

2. Material and methods

2.1. Speech data and human assessments

The data used in this study is a part of a larger speech corpus, which has been collected while piloting a computer-aided oral language test for large-scale evaluation [15]. The pilot sessions were conducted for groups of upper secondary school students in a classroom environment using headset microphones. Test tasks have a pool of trials, from which a random set is given to each examinee. For this study we selected utterances that occurred most frequently in the pilot test data, and for which the accurate placement of lexical and utterance stress was deemed important. The target utterances are listed, with their translations in Table 1. Twenty L2 productions of five read-aloud sentences were randomly selected for human assessments, and the

10.21437/Interspeech.2018-1873

Table 1: *Target utterances*

Utterance in Swedish	**English translation**
Allt fler högskolestudenter pluggar på distans i Sverige.	More and more highschool students are studying from distance in Sweden.
Bananer med droger i smugglades i tunnelbanan.	Bananas with drugs inside were smuggled in the underground.
Kyligt väder försenade jordgubbsskörden.	Chilly weather delayed the strawberry harvesting.
Dödsrisken 7,3 procent mindre bland cyklister med skyddshjälm.	Death risk 7,3 percent smaller among cyclists with safe helmet.
Recordmånga ålänningar gör frivilligt värnplikt.	Record number of Alanders volunteer for military service.

same utterances were extracted from five native speakers of the same Finland Swedish variant for reference. One native speaker was recorded in a studio setting; all other samples are from the pilot sessions.

Each speech sample was assessed by four native Finnish speaking teachers of Swedish, who were experienced in assessing spoken language skills and familiar with the rating scale used in this study. We used a six-level scale from the new CEFR descriptors for phonological control (levels A1, A2, B1, B2, C1, and C2, from the lowest to the highest) [16]. Here we focus on the assessments of prosodic features, which is a subsection of the CEFR scale for phonological control and pays attention to features such as word and sentence stress, rhythm and intonation with respect to the perceived intelligibility of speech.

The assessed speech data was manually annotated to syllable-level and f_0 was measured individually for each sample using the Praat program [17]. Additionally, four experts of Finland Swedish were asked to mark linguistically stressed syllables to the target utterances.

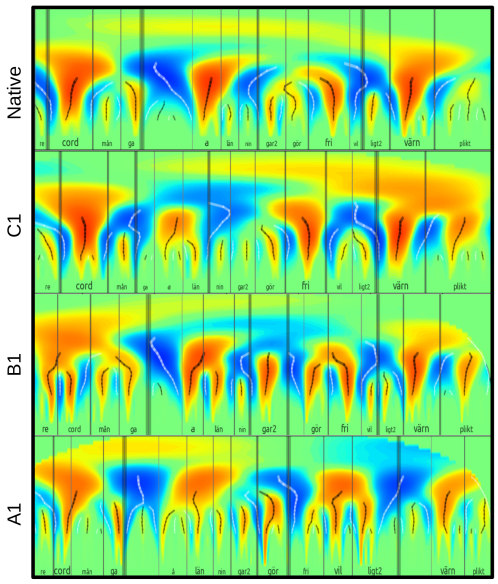

Figure 1: *Wavelet representation of prosody of a native utterance, compared to three L2 utterances in descending order of assessed proficiency. Warm colours indicate prominence, and lines of maximum amplitude in black represent the syllable prominence estimates.*

2.2. Wavelet-based prominence estimation

The prominence estimates for individual syllables were obtained using continuous wavelet analysis technique originally developed for word prominence detection described in [18]. The wavelet analysis was performed on seven combinations of fundamental frequency (f_0), energy envelope and duration signals: each signal separately, all three pairs of signals and a combination of all three signals.

First, f_0 and energy envelope signals were extracted and sampled at 200 Hz and then processed using methods described in detail in [18]. A duration signal was constructed in the following way: the value of each syllable duration was placed in the mid-time point of the unit and then connected using cubic interpolation to form a smooth duration signal.

Subsequently, the individual signals were then z-scored and different combinations were obtained by summing appropriate signals. Resulting combined signals were subjected to the continuous wavelet transform using a Mexican Hat mother wavelet, with scales a quarter of an octave apart. Lines of maximum amplitude were determined for each syllable from ten scales centered on average syllable length of the stimuli, yielding final prominence estimates (see Figure 1).

The prominence estimate methods are referred to by the signal combinations used as: f_0, DUR, EN, f_0-DUR, f_0-EN, EN-DUR and f_0-EN-DUR.

3. Results

3.1. Correlation among L1 speakers

As a measure of consistency between two renderings of the same sentence by two speakers we use a correlation between prominence estimates for corresponding syllables of the sentence. If the mutual consistency among the native speakers across all sentences is high enough, a mean native prominence estimates can be used as representation of an "average" native speaker producing a given sentence.

The correlations were calculated for each pair of L1 speakers and each sentence, separately for every prominence estimation method. The correlations for the same estimation method were pooled together.

Figure 2 depicts the pooled correlations for different methods as well as for all methods considered together. As can be seen the mutual correlations tend to be quite high, with medians of 0.78 (f_0), 0.77 (DUR), 0.51 (EN), 0.80 (f_0-DUR), 0.67 (f_0-EN), 0.64 (EN-DUR) and 0.74 (f_0-EN-DUR). Overall, the median correlation was 0.71.

The high overall correlations justify using the mean promi-

Table 2: *Summary of the best proportional odds logistic regression model for each prominence estimation method and two correlation types. The sign in min column indicates if the model with interaction (*) or without (+) was assessed as a minimal model.*

	Linguistic stress				**Average L1**			
	min	corr effect	t-value	AIC	min	corr effect	t-value	AIC
f_0	+	1.47	4.54	1214	+	1.72	5.53	1203
DUR	+	2.89	7.65	1172	+	3.27	9.26	1135
EN	+	2.50	7.88	1168	+	3.13	9.17	1143
f_0-DUR	+	2.39	7.04	1183	+	4.17	10.62	1101
f_0-EN	+	2.05	7.21	1180	*	4.24	6.80	1136
EN-DUR	+	3.27	9.90	1125	+	3.84	11.23	1083
f_0-EN-DUR	+	2.59	8.19	1162	+	3.68	10.73	1096

nence estimates for individual syllables of all five native speakers as a representation of L1 performance. We will use this "mean L1 prominence pattern" (one of for each sentence and estimation method) as a basis of evaluation of L2 speakers in terms of their realization of prominence.

3.2. Correlations of L1 with L2 and stress estimates

To compare the L2 and L1 renditions of the same sentences, the correlations between prominence estimates for L2 speakers and the average estimates for L1 were used; one correlation value for each graded L2 utterance and each method. These were calculated in the same way as the correlations among individual native speakers described above, on a sentence basis, separately for each estimation method. In addition, correlations between L2 prominence estimates and linguistic stress patterns (with 0 standing for unstressed syllables and 1 for the stressed ones) were calculated.

To answer the question whether correlation between L1 and L2 or stress patterns can be used to predict the proficiency level of individual L2 speakers, we used a proportional odds logistic regression model with grade (proficiency levels A1-C2 treated as an ordered factor) as a dependent variable and the correlation (either L2–L1 or L2–stress) and the assessor as independent variables. These models were fitted separately for each prominence estimate and each of the two correlation types. First, the minimal (simplest) model was sought using standard likelihood ratio test.

In general, there was no significant difference between the models with and without interaction between the independent

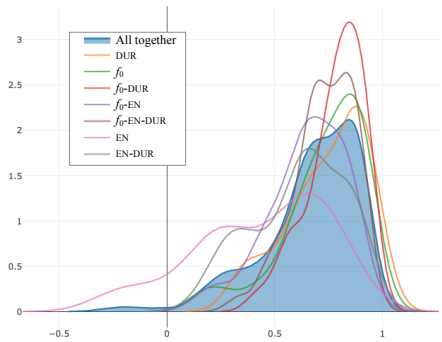

Figure 2: *The distribution of correlations between the native speaker prominence estimates.*

variable. (The only exception was the model with L1–L2 correlation obtained using f_0-EN estimate method, see Table 2.) In all cases, models with only one independent variable were significantly different from the models with both variables. Consequently, the models without interaction were selected as minimal models. For the f_0-EN, the model with interaction was chosen.

As seen in Table 2, for every model, the estimate of the correlation effect on grade is significantly positive, indicating that the greater agreement between L2 prominence pattern with L1 patterns as well as with linguistic stress leads to better grade, regardless to the prominence estimation method.

To compare different prominence estimation methods, Table 2 lists the Akaike information criterion (AIC) value for each model (the lower the AIC, the relatively better the predictive quality of fit). Note that for all estimation methods the values are lower when the grade is predicted using correlations between L2 and L1 performance than when correlation between L2 prominence patterns and linguistic stress is used. From all tested combinations, the correlations of L1 and L2 prominences estimated with energy envelope and duration (EN-DUR) yielded the lowest AIC (1083), i.e., the best quality of fit. Therefore, we have chosen this model to illustrate the nature of relationship between the prominence correlations and grades, separately for different assessors.

3.3. Evaluation of the best prediction

Figure 3 shows the distributions of correlations between L2 and (average) L1 prominence estimates for the EN-DUR model, separately for each assessor (in the left) and for all assessors pooled together (right). It clearly illustrates that, in general, higher correlation corresponds to better assessment grades. This seems to be the case particularly for the lower grades in the range A1–B1.

A *post hoc* Wilcoxon rank sum test (using appropriate Bonferroni correction for multiple comparisons) was used to compare correlation distributions for different grades (per assessor) supports this observation.

For the first assessor, the distributions for grades A1 and B1, A1 and B2, and A1 and C2 are significantly different ($p < 0.001$). For the second assessor, there are significant differences for grades A1–A2, A1–B2 ($p < 0.05$), A1–B1 ($p < 0.001$) and A1–C1 ($p < 0.01$), for the third for grades A1–B1 and A2–B1 ($p < 0.001$), and for the last assessor for grades A2–B2 ($p < 0.001$) and A2–C2 ($p < 0.01$).

When all correlation values for different assessors are pooled together, the distributions for A1 and A2 grades are both significantly different from those for all other grades ($p < 0.001$ for all pairs).

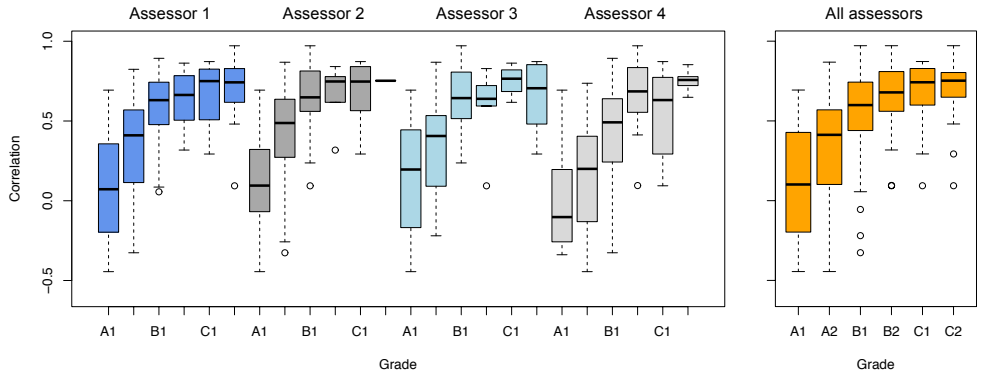

Figure 3: *The* EN-DUR *correlations and prosodic proficiency grades for each expert assessor.*

4. Discussion

This study investigated the effect of syllable-level prominence realization on the perceived proficiency level in Finnish learners of Swedish. Our results suggest that agreement between the native speakers and L2 speakers, as well as between L2 speakers and lexical stress patterns, correlates with the proficiency level (grade). This is the case regardless of the signal combination used for prominence estimation, although some combinations give better predictive power.

Interestingly, duration together with energy envelope presented the best signal combination for capturing syllable prominence characteristics correlating with L2 proficiency. On the other hand, adding f_0 signal to prominence estimator has in all cases a detrimental effect on the quality of fit. This is probably due to the fact that f_0 peaks do not necessarily align with the stressed syllable, but have other uses like boundary tones, at least in Finland Swedish. Therefore, f_0 might not serve as a reliable signal for prominence estimation, in particular on word-level. Actually, this result is in line with the findings of [19] that syllable level duration and energy serve as more reliable parameters for identifying L2 prosody, whereas f_0 provided poor identification results.

The L1 prominence estimates served as a more reliable predictor of L2 proficiency compared to binary, lexically determined stress patterns. This indicates, that our prominence estimate method captures not only word-level but also utterance-level prominence phenomena (lexical as well as sentence-level stress), and that correct sentence-level prosody contributes to perceived proficiency of L2 speakers.

As can be seen in Figure 1, the estimated syllable prominences (lines of maximum amplitude) differ between proficiency levels as well as from the native sample. Although the stress-related distinctions are generally maintained, the native and C1 samples show a greater difference between stressed and unstressed syllables. The prominence estimation method is designed to capture not only qualitative but also quantitative differences in the prominence level.

In the present context, the greater prominence difference between stressed and unstressed syllables may be associated with reduction in unstressed syllables that is prevalent in Swedish but much less pronounced in Finnish. This implies that especially less proficient L2 speakers tend to keep their native characteristics when speaking a foreign language. Detecting these characteristics can advance the development of automatic assessment systems.

The interaction between the correlations and assessor was mostly not significant for predicting the grade. This indicates that the assessors were consistent in their assessments in terms of grading more proficient L2 speakers better than less proficient ones, even if some assessor gave overall better grades than others.

The method proposed in this paper successfully captured the raters distinctions between less proficient (A-level) and fluent (C-level) students. Distinguishing between neighbouring proficiency levels, in particular at the higher end of the scale, was less reliable. As it is generally more challenging even for human raters to consistently differentiate between the levels of highly proficient speakers, it is possible that there exists a saturation effect, in particular in terms of syllable-level prominence. It must be also noted that the assessment distribution in our data is skewed towards lower proficiency levels with only about one in ten samples given a C-level assessment. The target sentences were elicited from a prototype test task that included tongue-twisting sentences that might be challenging even for native speakers to read. Therefore mistakes done in prominence production can be presumed to affect the intelligibility of speech and the perceived proficiency of the speaker.

Finally, the wavelet-based prominence estimation method has been shown to provide measurements potentially useful for assisted evaluation of L2 speaker proficiency. The several processing steps done manually for this investigation – segmentation, f_0 detection – can be automated given more formal examination setup. So, this method can serve as a part of an automated assessment system. As a part of integration process, we will assess to what extent the prominence based evaluation provides complementary information to the existing prosody evaluation measures such as disfluencies, pauses, and articulation rate [20].

5. Conclusions

This study shows that estimated syllable-level prominence patterns are a good predictor of L2 proficiency in terms of prosody.

Many studies on L2 prosody assessment and especially L2 prosody identification focus on f_0 features. Our results suggest, that in the assessment of L2 prosody, duration and energy cues could be even more significant than f_0.

6. Acknowledgements

The authors would like to thank Raili Hildén, Mikko Kuronen, Henna Heinonen and Huong Hoang for providing their expertise on Finland Swedish.

7. References

[1] M. J. Munro, T. M. Derwing, and S. L. Morton, "The mutual intelligibility of L2 speech," *Studies in second language acquisition*, vol. 28, no. 1, pp. 111–131, 2006.

[2] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL quarterly*, vol. 39, no. 3, pp. 399–423, 2005.

[3] L. D. Hahn, "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals," *TESOL quarterly*, vol. 38, no. 2, pp. 201–223, 2004.

[4] G. Fant and A. Kruckenberg, "Notes on stress and word accent in Swedish," in *Proceedings of the International Symposium on Prosody, Sept 18 1994, Yokohama*, 1994, pp. 2–3.

[5] G. Fant, A. Kruckenberg, and J. Liljencrants, "Acoustic-phonetic analysis of prominence in Swedish," in *Intonation*. Springer, 2000, pp. 55–86.

[6] K. Suomi and R. Ylitalo, "On durational correlates of word stress in Finnish," *Journal of Phonetics*, vol. 32, no. 1, pp. 35 – 63, 2004.

[7] M. Vainio and J. Järvikivi, "Tonal features, intensity, and word order in the perception of prominence," *Journal of Phonetics*, vol. 34, no. 3, pp. 319 – 342, 2006.

[8] H. Altmann and B. Kabak, "Second Language Phonology," in *Continuum Companion to Phonology*. Continuum, 2011, pp. 298–319.

[9] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[10] A. Wennerstrom, "The role of intonation in second language fluency," in *Perspectives on fluency*. University of Michigan, 2000, pp. 102–127.

[11] O. Engstrand and D. Krull, "Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of adaptive dispersion," *Phonetica*, vol. 51, no. 1-3, pp. 80–91, 1994.

[12] M. Kautonen, "Finskspråkiga talares intonation av finlandssvenska i påståendeyttranden i fritt tal," *Folkmålsstudier*, vol. 55, 2017.

[13] The Finnish Matriculation Examination Board: Kevään 2016 ja 2017 ylioppilastutkintoon ilmoittautuneiden määrät kokeittain. https://www.ylioppilastutkinto.fi/tietopalvelut/tilastotaulukot. Accessed: 2018-03-21.

[14] Gaudeamus igitur - ylioppilastutkinnon kehittäminen. Opetus- ja kulttuuriministeriön julkaisuja 2017:16. http://minedu.fi/julkaisu?pubid=URN:ISBN:978-952-263-462-7. Accessed: 2018-03-21.

[15] R. Karhila, A. Rouhe, P. Smit, A. Mansikkaniemi, H. Kallio, E. Lindroos, R. Hildén, M. Vainio, and M. Kurimo, "Digitala: An Augmented Test and Review Process Prototype for High-Stakes Spoken Foreign Language Examination." in *INTERSPEECH*, 2016, pp. 784–785.

[16] "Common European framework of reference for languages: Learning, teaching, assessment. Companion Volume with New Descriptors," Council of Europe, Education Department, Tech. Rep., 2017.

[17] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[18] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[19] M. Piat, D. Fohr, and I. Illina, "Foreign accent identification based on prosodic parameters," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[20] H. Kallio, J. Šimko, A. Huhta, R. Karhila, M. Vainio, E. Lindroos, R. Hildén, and M. Kurimo, "Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level," *Insights into second language speech*, no. 10, pp. 192–212, 2018.